

CLAIMS

What is claimed is:

1. A method of determining a similarity of a first string and a second string comprising:

5 calculating a Levenshtein matrix of said first string and said second string
determining a Levenshtein distance from said Levenshtein matrix; and
determining a largest common substring from said Levenshtein matrix.

2. The method according to Claim 1, wherein determining a largest common substring
10 from said Levenshtein distance matrix comprises determining a longest diagonal of equal
hamming distances of a lowest value.

3. The method according to Claim 1, further comprising calculating a Levenshtein
score.

15

4. The method according to Claim 1, further comprising determining the length of the
largest common substring.

5. The method according to Claim 4, further comprising calculating a largest common
20 substring score.

6. A method of determining a similarity of a first string and a second string comprising:

calculating a Levenshtein matrix of said first string and said second string;

determining a Levenshtein distance from said Levenshtein matrix;

5 determining a largest common substring from said Levenshtein distance matrix;

calculating a Levenshtein score as a function of said Levenshtein distance; and

calculating a largest common substring score as a function of said largest common substring.

10 7. The method according to Claim 6, further comprising calculating an acronym score.

8. The method according to Claim 7, further comprising calculating a weighted acronym score comprising a product of said acronym score and an acronym weight factor.

15 9. The method according to Claim 6, further comprising:

calculating a weighted Levenshtein score comprising a product of said Levenshtein score and a Levenshtein weight factor;

calculating a weighted largest common substring score comprising a product of said largest common substring score and a largest common substring weight factor; and

20 calculating a Levenshtein/largest common substring score comprising a sum of said weighted Levenshtein score and said weighted largest common substring score.

10. The method according to Claim 9, wherein a sum of said Levenshtein weight factor and said largest common substring weight factor is equal to one.

11. The method according to Claim 9, further comprising calculating a first weighted
5 numerical score comprising a product of said Levenshtein/largest common substring score and a string weight factor.

12. The method according to Claim 11, further comprising:
calculating an acronym score;
10 calculating a weighted acronym score comprising a product of said acronym score and an acronym weight factor; and
calculating a second weighted numerical score comprising a sum of said first weighted numerical score and said weighted acronym score.

13. The method according to Claim 12, wherein a sum of said string weight factor and said acronym weight factor is equal to one.

14. A computer-readable medium containing one or more sequences of instructions which when executed by a computing device cause the computing device to implement a
20 method for determining a similarity of a first string and a second string comprising:
calculating a Levenshtein score of said first string and said second string;

calculating a largest common substring score of said first string and said second string;
and
calculating a first numerical score as a function of said Levenshtein score and said
largest common substring score.

5

15. The computer-readable medium according to Claim 14, wherein calculating said
Levenshtein score comprises:

calculating a Levenshtein matrix of said first string and said second string;
determining a Levenshtein distance from said Levenshtein matrix; and
10 subtracting the resultant of dividing said Levenshtein distance by an average of a
length of said first string and a length of said second string from one.

16. The computer-readable medium according to Claim 14, wherein calculating said
largest common substring score comprises:

15 determining a length of a largest common substring from said Levenshtein matrix; and
dividing said length of said largest common substring by an average of a length of said
first string and a length of said second string.

17. The computer-readable medium according to Claim 14, wherein calculating said
20 first numerical score comprises:

calculating a weighted Levenshtein score comprising a product of said Levenshtein
score and a Levenshtein weight factor;

calculating a weighted largest common substring score comprising a product of said largest common substring score and a largest common substring weight factor; and
summing said weighted Levenshtein score and said weighted largest common substring score.

5

18. The computer-readable medium according to Claim 14, further comprising:
calculating an acronym score; and
calculating a second numerical score as a function of said first numerical score and said acronym score.

10

19. The computer-readable medium according to Claim 18, wherein calculating said second numerical score comprises:

calculating a weighted Levenshtein score comprising a product of said Levenshtein score and a Levenshtein weight factor;

15

calculating a weighted largest common substring score comprising a product of said largest common substring score and a largest common substring weight factor;

calculating a Levenshtein/largest common substring score comprising a sum of said weighted Levenshtein score and said weighted largest common substring score;

20

calculating a weighted Levenshtein/largest common substring score comprising a product of said Levenshtein/largest common substring score and a Levenshtein/largest common substring weight factor;

calculating a weighted acronym score comprising a product of said acronym score and an acronym score weight factor; and

summing said weighted Levenshtein/largest common substring score and said weighted acronym score.

5

20. The computer-readable medium according to Claim 19, further comprising:

utilizing said first numerical score for determining said similarity, when said first string and said second string comprise numerical-type strings; and

utilizing said second numerical score for determining said similarity, when said first

10 string or said second string comprise character-type strings.